DTIC FILE COPY

(2)

AFHRL-TP-88-6

AIR FORCE

HUMAN RESOURCES

AIR FORCE OFFICER QUALIFYING TEST (AFOQT): FORMS P PRE-IMPLEMENTATION ANALYSES AND EQUATING

Kurt W. Steuck
Thomas W. Watson
Jacobina Skinner

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

November 1988
Interim Technical Paper for Period January 1986 - January 1988

Approved for public release; distribution is unlimited.

LABORATORY

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) AFHRL-TP-88-6	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Manpower and Personnel Division 6b. OFFICE SYMBOL (If applicable) AFHRL/MOAO	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) Air Force Human Resources Laboratory Brooks Air Force Base, Texas 78235-5601	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Air Force Human Resources Laboratory 8b. OFFICE SYMBOL (If applicable) HQ AFHRL	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code) Brooks Air Force Base, Texas 78235-5601	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 62703F; PROJECT NO. 7719; TASK NO. 18; WORK UNIT ACCESSION NO 47

11. TITLE (Include Security Classification)
Air Force Officer Qualifying Test (AFOQT): Forms P Pre-Implementation Analyses and Equating

12. PERSONAL AUTHOR(S)
Steuck, K.W.; Watson, T.W.; Skinner, J.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Interim	FROM Jan 86 TO Jan 88	November 1988	48

16. SUPPLEMENTARY NOTATION
Analyses were conducted in AFHRL/TS Study numbers 9182 and 9574.

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Air Force Officer Qualifying Test (AFOQT) item analyses,
05	08		aptitude tests, selection tests.
05	09		

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Forms P of the Air Force Officer Qualifying Test (AFOQT) became operational in 1987. This paper describes the equating of Forms P_1 and P_2 to Form 0, and the associated analyses.

The pre-implementation equating was necessary (a) to check the adequacy of the items in the new forms, (b) to assess the similarity of the new forms, and (c) to establish conversion tables for placing scores from the new test on the metric of Form 0. Three forms of the AFOQT (0, P_1, and P_2) were administered to subjects (N = 3,400) at Basic Military Training School (BMTS), Air Force Reserve Officer Training Corps (AFROTC), and Officer Training School (OTS). Results are reported at the item, subtest, and composite levels. In general, Forms 0 and P are content equivalent while Forms P have items of slightly lower difficulty than those of Form 0. Equipercentile equatings were used to produce provisional conversion tables for the operational implementation of Form P. Keywords:

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Nancy J. Allin, Chief, STINFO Branch	22b. TELEPHONE (Include Area Code) (512) 536-3877 / 22c. OFFICE SYMBOL AFHRL/SCV

DD Form 1473, JUN 86 — Previous editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

AFHRL Technical Paper 88-6 November 1988

AIR FORCE OFFICER QUALIFYING TEST (AFOQT): FORMS P PRE-IMPLEMENTATION ANALYSES AND EQUATING

Kurt W. Steuck
Thomas W. Watson
Jacobina Skinner

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

Reviewed by

David E. Brown, Lt Col, USAF
Chief, Officer Selection and Classification Function

Submitted for publication by

Lonnie D. Valentine, Jr.
Chief, Force Acquisition Branch

This publication is primarily a working paper. It is published solely to document work performed.

SUMMARY

This paper describes the provisional equating of Forms P_1 and P_2 of the Air Force Officer Qualifying Test (AFOQT) and the associated analyses in preparation for its operational implementation in 1987. The pre-implementation equating was necessary (a) to check the adequacy of the items in the new forms, (b) to assess the similarity of the new forms, and (c) to establish conversion tables for placing scores from the new test on the metric of Form O. Three forms of the AFOQT (O, P_1, and P_2) were administered to about 3,400 military subjects at 11 Air Force bases. The subjects were from Basic Military Training School (BMTS), Air Force Reserve Officer Training Corps (AFROTC), and Officer Training School (OTS).

Analyses were computed at the item, subtest, and composite levels and several types of equatings were completed. The distributions of items based on item difficulty and item discrimination were similar across forms, but not identical. Equipercentile equatings were used to produce conversion tables for Forms P for provisional use prior to the Initial Operational Test and Evaluation (IOT&E) of the new forms.

PREFACE

The Air Force Human Resources Laboratory (AFHRL) is tasked as the test development agency for the Air Force Officer Qualifying Test (AFOQT) by Air Force Regulation 35-8, Air Force Military Personnel Testing System. The current research and development (R&D) effort was undertaken as part of AFHRL's responsibility to develop, revise, and conduct research in support of the AFOQT. Work was accomplished under Task 771918, Selection and Classification Technologies, which is part of a larger effort in Force Acquisition and Distribution Systems. The study was completed under Work Units 77191847 (Development and Validation of Civilian and Nonrated Officer Selection Methodologies) and 77191824 (Officer Item Pool Development).

The authors would like to thank their colleagues in the Manpower and Personnel Division for their assistance in this effort. Mr. Todd Sperl provided adroit assistance with data tabulation and analysis, and Dr. Malcolm Ree provided expert advice on a variety of technical issues. A number of colleagues supported this effort by going on-site to serve as test administrators or proctors. Specifically, we extend our appreciation to 1Lt Thomas O. Arth, Mr. Roy E. Chollman, Mr. Douglas K. Cowan, Mr. Refugio Gonzalez, Jr., and A1C Bertrand L. Washer.

The authors acknowledge with considerable gratitude the assistance of Ms. Doris E. Black, Mr. James L. Friemann, A1C Dave Lawson, Sgt Dave LeBrun, and Ms. Suzanne Farrell of the Information Sciences Division, AFHRL. Their efforts were instrumental to the successful accomplishment of the data analysis phase of this study.

Thanks are also expressed to the many operational managers and training staff members associated with the Air Staff, the Air Force Military Personnel Center (AFMPC), Air Training Command (ATC), Air Force Reserve Officer Training Corps (AFROTC), Basic Military Training Center (BMTC), and Officer Training School (OTS). Managers in these organizations made it possible for the testing to take place, despite inconvenience to ongoing training which was often considerable. Numerous training staff personnel throughout the Continental United States (CONUS) provided the on-site assistance which was essential to the successful data collection. When necessary, they even assisted AFHRL staff with proctoring. We also appreciate the assistance of the thousands of cadets and basic military trainees who took the various forms of the AFOQT.

Finally, we wish to thank the staff of Psychometrics, Inc., especially Drs. Ray and Frances Berger, and Dr. Willa Gupta, who did such an excellent job in developing the best possible AFOQT Forms P under the constraint that the forms be parallel in content and format to Form O.

TABLE OF CONTENTS

LIST OF TABLES

AIR FORCE OFFICER QUALIFYING TEST (AFOQT):
FORMS P PRE-IMPLEMENTATION ANALYSES AND EQUATING

I. INTRODUCTION

Background of the Air Force Officer Qualifying Test (AFOQT)

The United States Air Force currently selects officers from three applicant pools. One pool consists of highly qualified high school graduates who are accepted on the basis of Congressional recommendations and other criteria into the United States Air Force Academy (USAFA) at Colorado Springs, Colorado. After completing a 4-year college program, graduates enter the Air Force as second lieutenants. Since the Scholastic Aptitude Test is used as the primary selection tool for these individuals, they are not required to take the Air Force Officer Qualifying Test (AFOQT) for selection purposes. The second pool of applicants enters the Air Force through the Air Force Reserve Officer Training Corps (AFROTC). These individuals attend universities and colleges throughout the nation and enroll in AFROTC courses in their last 2 years of schooling. The majority take the AFOQT as high school seniors or before their junior year of college. The third pool of applicants consists of men and women who have completed a baccalaureate degree at an accredited university or college and apply for Officer Training School (OTS). These individuals take the AFOQT for selection into OTS and are commissioned upon completion of the OTS program.

Although the selection of aircrew members dates back to World War I, the first screening test for preliminary selection of officers, the Aviation Cadet Qualifying Examination, was published in 1942. Various iterations of the original test, with different names, were used for selection screening over the next decade, complemented by the Aircrew Classification Battery (ACB). These instruments underwent considerable change during this era of experimentation. By 1952, a preliminary version of the AFOQT was developed and by 1955, the AFOQT replaced the ACB and its screening test predecessors. Since that time, the AFOQT has been updated periodically. Although items have changed and subtests have been added or deleted, the composite structure of the AFOQT has remained rather constant through the years. The experimentation of the 1940s and early 1950s gave way to evolutionary refinement in more recent decades. Interested readers should consult Rogers, Roach, and Short (1986) for information about the selection of commissioned officers and a brief history of testing of Air Force officers.

Recent forms of the AFOQT have been dramatically shortened, and the subtest structure has been modified. Form N of the AFOQT, implemented in 1978, consisted of 606 items divided into 18 subtests (Gould, 1978). These subtests were used to compute the following five composite scores: Pilot, Navigator-Technical, Officer Quality, Verbal, and Quantitative. In contrast, with operational implementation of AFOQT Form O in 1981, substantial changes in content, format, administration, and scoring were made (see Rogers, Roach, & Wegner, 1986 for details). Form O consists of 380 items (226 fewer than Form N) and is divided into 16 rather than 18 subtests. Although the composites are similar (Officer Quality was renamed Academic Aptitude), four subtests were dropped and two new ones were added. Furthermore, the amount of time required for administration was reduced from about 7 hours to 4.5 hours. Table 1 shows the number of items in each subtest and how the subtests are arranged into the five composites for Forms N and O. Because the number of items and subtest/composite structure of Forms P (discussed in the section which follows) are so similar to previous forms, the composition of Forms P is also shown in Table 1.

Table 1. Item, Subtest, and Composite Structure for AFOQT Forms N, O, and P

Subtest	No. of Items in N	No. of Items in O and P	Pilot	Navigator-Technical	Academic Aptitude	Verbal	Quantitative
Verbal Analogies	25	25	A[a]		A	A	
Arithmetic Reasoning	25	25		A	A		A
Reading Comprehension	25	25			A	A	
Data Interpretation	25	25		A	A		A
Word Knowledge	25	25			A	A	
Math Knowledge	25	25		A	A		A
Mechanical Comprehension	24	20	A	A			
Electrical Maze	30	20	A	A			
Scale Reading	48	40	A	A			
Instrument Comprehension	24	20	A				
Block Counting	80	20	A	A			
Table Reading	50	40	A	A			
Aviation Information		20	O P[b]				
Rotated Blocks	20	15		A			
General Science	24	20		A			
Hidden Figures		15		O P			
Background for Current Events	25		N[c]		N	N	
Tools	25		N	N			
Aerial Landmarks	40			N			
Pilot Biographic and Attitude Scale	66		N				
Total	606	380					

[a]A = N, O, and P.
[b]O P = O and P.
[c]N = N Only.

Development of AFOQT Forms P

Historically, the Air Force Human Resources Laboratory (AFHRL) has been responsible for periodic updates of the AFOQT, including the new Forms P_1 and P_2 which became operational in June 1987. In support of this responsibility, AFHRL contracted with Psychometrics, Inc., of Sherman Oaks, California, to develop a large pool of items in content areas already covered by the AFOQT. From the extensive pool of items developed in each of the existing content areas, AFHRL and Psychometrics, Inc. selected items to be used in conjunction with existing items from previous forms to create two parallel versions of Form P.

Those previous items which are common across Form O and Forms P are referred to as anchor items. They link the three forms for experimental purposes, and were selected on the basis of their performance with officer applicant samples. Empirical data also provided the basis for developing and selecting new items for AFOQT Forms P. New items were combined with anchor items in several sets of experimental booklets and administered primarily to airmen in Basic Military Training School (BMTS). In some instances, especially for difficult subtests, experimental booklets were also administered to OTS cadets. Items were evaluated using classical item analyses. Officer item difficulty estimates were generated to supplement actual difficulty indices obtained from the airmen samples. New items meeting a variety of psychometric criteria for difficulty, discrimination, and content were selected for inclusion in the new AFOQT Forms P.

Rationale for the Current Investigation

Despite the extensive research performed in the development of AFOQT Forms P, the current investigation was a necessary adjunct. The adequacy of the items which comprise Forms P had already been assessed, but needed to be checked using data from a more representative sample composed primarily of officer candidates rather than airmen trainees. In addition, the new Forms P had to be compared, not only with each other but with Form O, to determine how parallel they were, and equating analyses had to be conducted to provide conversion tables linking scores on the new forms with those on the previous form of the AFOQT. Thus, the goals of this investigation are threefold: (a) to verify the adequacy of items in Forms P using a more representative sample; (b) to compare Forms O, P_1, and P_2 to determine if they are parallel, as designed; and (c) to derive scores on Forms P that are comparable to scores on Form O.

Determining the Adequacy of Items in Forms P Using a More Representative Sample

Test construction procedures used to develop Forms P were designed to identify items which were psychometrically sound. However, as indicated previously, the primary empirical basis for judgments concerning the adequacy of items was analyses performed on data obtained from airmen samples. Reliance on data from airmen subjects in the development of items for use with officer applicants is not ideal. Obviously, the sample on which items were developed is not representative of the target population to be tested. Differences in age, education, and aptitude may limit the generalizability of the data obtained.

Considering the drawbacks of using airmen samples, the rationale for their use to screen candidate items for the AFOQT needs to be explained. A huge item pool was developed which needed to be administered to a larger number of subjects than was available (without considerable time and expense) from the pool of officer candidates. For each of the 16 content areas in Forms O and P, 300 new items were developed. Each of these items was administered to at least 350 subjects. Considering the magnitude of the item development task and the limited supply of officer candidates, the use of airmen samples, augmented only occasionally with officer candidates, was a logistical and economic necessity. However, it was also necessary to confirm,

prior to operational implementation, that the items selected for Forms P performed well when tested on a more representative sample. If they did not, then adjustments could be made in the test prior to final printing and operational use.

Comparing Forms O, P_1, and P_2 to Determine if They Are Parallel

A design goal in developing the two new versions of Form P was to construct parallel tests which were also parallel to Form O. Briefly, the general procedure used was to match psychometric characteristics (based largely on airmen samples) of items occupying the same position on Forms O, P_1, and P_2. A second objective of the current investigation was to determine the actual degree of parallelism among the three forms based primarily on officer samples.

Deriving Scores on Forms P That Are Comparable to Scores on Form O

As discussed above, parallelism in tests is a design goal which can be attained only imperfectly. Thus, despite the parallel design of the forms, scores on Forms P_1 or P_2 would not be exactly equivalent to the same score on previous forms without further equating. However, as Angoff (1971) has discussed, techniques exist which allow scores derived from different forms, after conversion, to be directly equivalent. Thus, the third objective of this investigation was to perform equating analyses which would provide an empirical basis for generating two separate sets of provisional conversion tables (one set for Form P_1, the other set for Form P_2) to link scores on these tests to scores on Form O. These provisional conversion tables would be modified, if necessary, based on the results of the Initial Operational Test and Evaluation (IOT&E).

II. METHOD

Subjects

Rationale for Subject Selection

Subjects were 3,376 airmen and officer students in BMT, OTS, and AFROTC who were administered either Form O, Form P_1 or Form P_2 of the AFOQT. The total number of examinees by training program and AFOQT form is shown in Table 2. The total sample size was reduced, following data cleanup, to a computational sample of 3,341 cases. Analyses conducted by test form were based on the following case counts: $N = 1{,}101$ for Form O, $N = 1{,}120$ for Form P_1, and $N = 1{,}120$ for Form P_2. Subjects were all students in these training facilities who were available for testing from 31 May 1986 to 18 July 1986. The timeframe was limited by the need to prepare conversion tables in time for the target operational implementation date and to detect and correct problems, if any, in the Forms P booklets prior to final printing. Start and stop dates were based on practical considerations. The stop date allowed collection of data from a desired minimum of 3,000 subjects, and also provided sufficient time for data analysis and interpretation, and any final modification to the Forms P booklets prior to the printing deadline. Camera-ready copies of the final forms were due on 1 October 1986 at the Air Force Military Personnel Center (AFMPC), the agency responsible for administrative oversight of the operational testing program.

Table 2. **Number of Examinees by Training Program and AFOQT Form**[a]

	Form			
	0	P_1	P_2	Total
BMT Airmen	258	256	255	769
	(23)	(23)	(23)	(23)
OTS Cadet	194	194	195	583
	(17)	(17)	(17)	(17)
AFROTC Cadet	642	667	666	1,975
	(58)	(59)	(59)	(59)
Unknown	17	16	16	49
	(2)	(1)	(1)	(1)
Total N	1,111	1,133	1,132	3,376
	(100)	(100)	(100)	(100)

[a]Cell values shown in parentheses below Ns are percentages of the total column frequency.

The rationale for subject selection needs to be elaborated. For the pre-implementation evaluation described in this paper, a major goal was to obtain data from sufficient subjects at all points throughout the range of abilities (or performances, as measured by test scores) to generate preliminary conversion tables. Three groups (i.e., BMTS students, AFROTC cadets, and OTS cadets) were selected for participation since they were expected to score, on the average, at different points along the score continuum for the various subtests or composites. Due to age and educational level, BMT airmen were expected to provide scores primarily at the lower end of the continua, whereas AFROTC subjects, who were still in school, were expected to "fill in" the middle range. OTS subjects, who had completed their baccalaureate degrees, were expected to score at the higher ranges.

Procedures for Subject and Site Selection

Both BMTS and OTS are located at Lackland Air Force Base, San Antonio, Texas, which is also the site of the AFHRL testing facility. AFROTC facilities are scattered at many colleges and universities throughout the country. However, during the data collection period, AFROTC subjects were temporarily assigned to 11 field training sites. This permitted representative sampling of AFROTC students at considerable savings of time and travel expenses. The sites involved, and the numbers and types of subjects tested, are provided in Table 3.

Demographic Characteristics of Subjects

The following description of demographic characteristics is based on the computational sample of 3,341 subjects.[1] Most subjects were males (2,689 or 81%); 645 or 19% were females. Most were white (2,808 or 84%) while 286 (9%) were black, and 7% of other ethnic origin. Ages ranged from 17 years to 34 years, with the majority (76%) being 22 years of age or younger. Education ranged from 12 to 21 years, with most subjects having had some college. Only 15% (n = 514) had 12 years of education, while 80% had between 13 years and 16 years of education. Educational credentials ranged from high school diplomas to masters' degrees. However, 793 or 24% had an associate or baccalaureate degree. Only 1% had been awarded a master's degree.

[1]Due to missing demographic data on some cases, the frequencies may not sum to 3,341.

Table 3. Distribution of Examinee Categories by Testing Site[a]

	Test Form		
	0	P_1	P_2
Basic Airmen			
Lackland AFB (LAFB)	258	256	255
	(23)	(23)	(23)
Officer Training School Cadets			
Medina Annex of LAFB	194	194	195
	(17)	(17)	(17)
AFROTC Cadets			
McChord AFB	59	64	63
	(5)	(6)	(6)
McClellan AFB	39	44	37
	(4)	(4)	(3)
Tyndall AFB	28	29	29
	(3)	(3)	(3)
Robins AFB	38	38	37
	(3)	(3)	(3)
Dover AFB	37	37	36
	(3)	(3)	(3)
Wright-Patterson AFB	31	37	37
	(3)	(3)	(3)
Plattsburgh AFB	60	62	61
	(5)	(5)	(5)
McConnell AFB	69	71	72
	(6)	(6)	(6)
Vandenberg AFB	102	100	100
	(9)	(9)	(9)
Bergstrom AFB	29	32	31
	(3)	(3)	(3)
Lackland AFB	150	153	163
	(14)	(14)	(14)
Unknown	17	16	16
	(2)	(1)	(1)
Total N	1,111	1,133	1,132
	(100)	(100)	(100)

[a]Cell values shown in parentheses below Ns are percentages of the total column frequency.

Administrative Procedures

Testing at AFROTC Field Training Sites

A testing schedule was arranged that would allow two-person teams from AFHRL to make five trips of 3 to 5 days' duration, usually to multiple sites. These trips were scheduled sequentially to ensure that sufficient materials would be available for administration. Team composition varied from trip to trip. The AFOQT forms were administered on days available in the field-site training schedule, including Saturday and Sunday. Total testing time was approximately 4 1/2 hours, excluding initial preparation and clean-up. Once testing was completed at one site, AFHRL teams typically had at least a day to travel to the next site, make

final arrangements, and orient on-site personnel who assisted with proctoring. Wherever possible, single morning and/or afternoon sessions were conducted with the AFHRL team serving as test administrator and lead proctor. Large rooms, such as a large testing room, the ballroom of an officer's club, or a recreation center were used. Although there were some unavoidable variations in the quality and configuration of facilities from site to site, care was taken to ensure that the test administration environment was as standardized as possible and adequate for administration of the AFOQT. In a few instances, two administration sessions were conducted simultaneously, with both AFHRL team members serving as test administrators, assisted by on-site proctors.

Testing at BMT and OTS Sites

Since BMT and OTS training facilities are collocated at Lackland AFB with the AFHRL testing facility and staff, arrangements for testing at these sites did not involve extensive travel and required no proctoring assistance from the training staffs of these schools. BMT subjects were tested in their own facilities, the AFHRL facilities, or some other suitable facility at Lackland. OTS subjects were tested in the OTS auditorium. Due to the unavailability of 4 1/2- to 5-hour periods of time in their training schedule, OTS subjects were administered the AFOQT in two 2 1/2-hour segments. This deviation from the procedures used with the other groups was unavoidable. In other respects, testing procedures were as similar as possible to those used with AFROTC subjects, except that AFHRL staff performed all administrator and proctor duties.

Development of a Manual for Administration

To ensure effective and consistent administration of the multiple AFOQT forms during this investigation, a separate Manual for Administration was prepared. The existing Form O Manual for Administration was being revised in preparation for development of a Forms P Manual for Administration. The Form O version was adapted for operational use in the current investigation. Changes were made so administrators would alert examinees to the need to identify color-coded test booklets as one of the three forms on the answer sheet, and, in the case of Forms P, to identify the correct version number. Whenever necessary, changes were also made (a) to identify across-forms differences in the wording of directions or other attributes of the tests or their administration, (b) to reflect the experimental nature of the testing, and (c) to alert administrators to read a separate Privacy Act Statement suitable to the experimental nature of the testing sessions.

Selection and Training of Administrators and Proctors

To ensure that test administration procedures were as standardized as possible, considerable emphasis was placed on the selection and training of administrators and proctors. This was especially important since the test was being administered by multiple administrators/proctors in multiple settings. Test administrators and proctors were either psychologists with a good understanding of psychometric principles or experienced test administrators. Nearly all had prior test administration experience. Those with the most experience, regardless of rank, were assigned as test administrators in the administrator/proctor teams.

Training materials were developed by AFHRL scientists, and a 1/2-day training session was held in which AFHRL staff members responsible for administration or proctoring participated. Topics involved AFOQT test administration, orientation and training of on-site personnel, setting up of the testing room, distribution of test materials, safeguarding of the tests and answer sheets, and preliminary data checks.

To help ensure that on-site AFROTC personnel were prepared to assist with proctoring, Air Force regulations relevant to testing were forwarded in advance to the field sites, along with a description of the specific duties of test proctors. In addition, AFHRL test administrators were instructed to review proctoring duties with proctors upon arrival at the site, and AFHRL team members were requested to supervise on-site proctors during the testing sessions.

Use of Two Different Answer Sheets

Prior to each testing session, answer sheets were placed inside the front cover of each test booklet. Two types of answer sheets were used: one red, the other green, each differing slightly in format and in the size and shape of the response ovals or bubbles. The rationale for use of two different answer sheets requires explication. The red answer sheet has been used operationally in the administration of the AFOQT Form O to AFROTC cadets. The green answer sheet has been used in the operational administration of AFOQT Form O to OTS applicants. Two different answer sheets had been used in obtaining Form O data since scoring had been decentralized (at Maxwell AFB and either Brooks AFB or Randolph AFB), and the scanning equipment at Maxwell AFB was unable to process the green sheets. The nature of the answer sheet can affect test scores, especially on speeded subtests, as pointed out by Wegner and Ree (1985).[2]

In collecting Form O data, it was necessary to use both answer sheets to approximate the previous operational practice. Thus, in preparation for administration at each site, green answer sheets were inserted in Form O test booklets to be administered to OTS cadets and red answer sheets in test booklets to be administered to AFROTC cadets. The two types of answer sheets were equally divided among the BMTS examinees. Forms P booklets were prepared for administration by inserting only green answer sheets, consistent with the new operational use of a single type of answer sheet.

Administration of AFOQT Forms O and P

Prior to entry of the examinees into each testing session, test administrators and proctors placed a copy of each form (with its answer sheet insert) in sequential order (O, P_1, P_2) at each testing station. This was done to ensure that randomly equivalent groups were formed. If multiple testing sessions occurred at a site, distribution of the booklets was counterbalanced by starting with the next booklet in the series on simultaneous or subsequent testing sessions. Thus, if the last booklet distributed in one session was a Form P_1 booklet, the first booklet distributed at the next testing session was a Form P_2 booklet.

Data Analyses

Scoring and Data Editing

In order to provide accurate scoring of the optical scan answer sheets, they were first checked by test administrators to determine if they were suitable for scanning and to correct problems such as removing stray marks, darkening ovals, etc. Green answer sheets were then scanned by the Technical Services Division (TS)[3] of AFHRL, while red answer sheets were forwarded to HQ AFROTC for scanning.

---

[2]To avoid such potential for error in future operational use, scoring will be centralized at AFMPC at Randolph AFB and the green answer sheet will be the only one used with the operational implementation of Forms P.

[3]Division has been redesignated the Information Sciences Division.

Power - Speed Issue

Historically, the AFOQT has contained subtests which have been described as conforming to speeded, power, or mixed models. A power model is one in which all examinees have enough time to consider every question in the subtest. A speeded test is one in which the items are easy, but there is not enough time for each individual to answer every item (Gulliksen, 1950). Therefore, successive items are reached by fewer and fewer examinees and individuals responding at a slower rate have time to consider fewer items than those responding at a faster rate. Mixed models are those that have items written as in a power test, but yet examinees do not have enough time to answer every item. Mixed models, then, do not follow either a pure power or pure speeded model.

The index used to evaluate whether a given subtest conforms to a power, speed, or mixed model is the proportion of subjects not responding to items in the subtest (i.e., proportion of omitting). Power tests characteristically have items with proportions of omitting that are less than .05. When the item number is plotted with the corresponding proportion of omitting, power tests exhibit a flat line. True speeded subtests have low omitting rates for items at the beginning of the subtest, but then show a steady increase in omitting rates for items in the last half of the subtest. Mixed model subtests are defined as having low, flat rates for the majority of the items, but have an increase in omitting rate for the final few items.

Whether a subtest is best described as power or speeded has two implications. One implication concerns the interpretation of test results. Not having an appropriate understanding of the nature of the subtest could lead to misinterpretation of an examinee's knowledge and abilities. A second implication is that knowledge of degree of speededness should guide decisions about data analysis. For example, the computational formulas for item difficulty and item discrimination indices differ for power and speeded tests. If an inappropriate analysis is used, the item statistics computed may underestimate or overestimate the "true" statistics.

Skinner and Ree (1987) categorized the 16 AFOQT subtests according to the model to which each conformed. Mechanical Comprehension, Rotated Blocks, and General Science were judged to be power subtests; Electrical Maze, Instrument Comprehension, and Block Counting were classified as speeded; and the remaining subtests were described as following a mixed model. This pattern is similar but not identical to the classifications made by Gould (1978) and Miller (1974). These two studies designated Electrical Maze, Instrument Comprehension, Scale Reading, Table Reading, and Block Counting as being speeded. Nonetheless, it should be noted that, in the Skinner and Ree data, these five subtests have highly speeded components. Later in this paper, the degree of speededness for the subtests in Forms 0, P_1, and P_2 will be discussed.

Classical Item Analysis

The analysis of performance of Forms P_1 and P_2 at the item level was based on classical or "true score" theory (Gulliksen, 1950; Henrysson, 1971). Item difficulties ($\underline{p}$) were calculated as the proportion of examinees responding correctly to the item. The biserial correlation (r_{bis}) between the item score (correct or incorrect) and total subtest score was used as the index of the discrimination value of each item. In the analysis of power subtests, the level of difficulty is calculated by dividing the number of examinees selecting the correct option by the number of examinees taking the subtest. Since for power subtests it is assumed that all items are reached by all examinees, the number of people attempting each item equals the number of examinees taking the subtest. In contrast, in the analysis of speeded subtests, the total number of examinees taking the test is not used in calculating the level of difficulty. Rather, difficulty is defined as the number of examinees selecting the correct option divided by the number of examinees who make a response to the item or one later in the subtest. Examinees who do not finish a subtest are not included in the analyses of the items they do not reach.

Subtest and Composite Analysis

Subtest and composite raw scores were described and compared in terms of their mean, standard deviation, skewness, kurtosis, and reliability. Proportion correct was also obtained. Test reliability was computed using Kuder-Richardson Formula 20 for power subtests. Intercorrelations were also calculated for subtest and composite raw scores.

Equating Design

An equivalent, random groups design (Angoff, 1971) was used to equate Forms P_1 and P_2 to Form O. Forms P_1 and P_2 were constructed to be parallel to each other and Form O in content, difficulty, and reliability. Because the new forms were content parallel, an equating, as opposed to a calibration, was conducted.

The approach taken to equate the new forms gave consideration to the use of different answer sheets in the operational AFOQT Form O testing program. An inspection of testing load frequencies indicated that the number of officer applicants examined each year on the two answer sheets was roughly equal. That is, about half of the examinees were OTS applicants tested on the green answer sheet and the other half were AFROTC applicants tested on the red answer sheet. The proportion of examinees by answer sheet type observed in the operational program needed to be preserved in the equating analyses to account for potential effects that the different answer sheet features may have had on Form O performance. The answer sheets differed not only in color but also in structural features such as response grid and arrangement. A weighting procedure was devised to obtain a Form O distribution for equating analyses which gave equal weight to the scores of subjects in the current sample tested on either answer sheet. As described in a previous section of this paper, fewer of the Form O subjects had been supplied a green answer sheet than a red answer sheet. Therefore, scores for the remaining subjects were weighted by 2.661 (the ratio of number of examinees tested on red sheets to those tested on green sheets) to yield a Form O distribution in which the scores obtained from the two answer sheet types were represented in equal numbers.

Linear and equipercentile equatings were accomplished as described by Angoff (1971, pp. 568-573) for each composite of each form of the new test. In the linear method, on equivalent forms raw scores that have the same z-score value are set equivalent; in the equipercentile method, raw scores that have the same percentile rank are set equivalent. Since the equipercentile method may produce irregular equating curves, three forms of smoothing were conducted. Linear, quadratic, and cubic polynomial regressions for smoothing were used with the Form P scores entered as the independent variables and the Form O scores serving as the dependent variables. For linear smoothing the first power of the independent variable was entered as the independent variable into a multiple regression equation; for quadratic, the first and second powers were entered; and for cubic, the first, second, and third powers were entered. As a result, four equatings (i.e., one linear equating and three smoothings of equipercentile equating) were produced for each composite on Form P_1 and four for each composite on P_2.

Decisions had to be made about which method was appropriate for each composite. The decisions were based on the similarity of the distribution of the equated new test scores (e.g., Form P_1) with the distribution of the reference test scores (i.e., Form O) (Braun & Holland, 1982). The method of equating which produced the greatest similarity in the two distributions was selected. Several statistical indices of goodness-of-fit were examined to distinguish among the equatings. These were (a) the standard error of estimate for polynomial smoothing techniques and (b) three measures of deviation between raw scores (bias, absolute average deviation, and root mean square deviation) for equipercentile versus linear (z-score) equatings.

III. RESULTS AND DISCUSSION

Twelve items were previously removed from Form 0 due to double keys, miskeys, or poor item performance. Three items were removed from Verbal Analogies, four items from Arithmetic Reasoning, two from Data Interpretation, one from Word Knowledge, one from Mechanical Comprehension, and one from Scale Reading. Because these items were removed from the current analyses, the number of items per subtest for Form 0 differs from those for the corresponding subtests in Forms P_1 and P_2. This influences the comparison of Form 0 with Forms P_1 and P_2, but not the comparison of P_1 with P_2.

The item omitting rates for Forms 0, P_1, and P_2 in these samples were used to determine the type of analysis (power, speeded, or mixed) for each subtest. While it is appropriate to make these determinations for analysis purposes, no permanent reclassification of the subtests as speeded or non-speeded is implied. The patterns of omitting in the samples were consistent across the three forms for each subtest. Mechanical Comprehension, Rotated Blocks, and General Science conformed to the power model, whereas Electrical Maze, Scale Reading, Instrument Comprehension, Block Counting, and Table Reading exhibited highly speeded components. The remaining eight subtests showed slightly speeded or mixed-model components. Therefore, it was decided to analyze the five highly speeded subtests using the speeded computational formulae for item difficulty and discrimination, while analyzing the remaining subtests, even if slightly speeded, as power subtests.

In discussing the results of this investigation, general comments will be provided first. Then each of the three forms will be discussed individually in the following order: Form 0, Form P_1, and Form P_2. Next, comparisons will be made between Form 0 and Forms P_1 and P_2. Finally, comparisons will be made between Forms P_1 and P_2.

Item Analysis

Item Difficulty. Item difficulty is based on the proportion of individuals selecting the correct option for a given question and is dependent on the ability in the sample (sample specificity). Difficulties have a range of .00 to 1.00. Items with values between .00 and .30 have a low proportion of people selecting the correct option and therefore are considered to be hard. Items with values between .70 and 1.00 have a high proportion of people selecting the correct option and therefore are considered easy. Interpretation can be confusing in that an item with a high item difficulty index (e.g., .90) is an easy item. The converse is that an item with a low difficulty index (e.g., .20) is a difficult item.

The reader should note that the following discussion must be interpreted with care due to the way in which the item difficulty values were distributed. The categories reported in Table 4 are arbitrary and other categories could have been generated (e.g., .31 to .50), resulting in slightly different summarizations. Furthermore, scores within a category may be farther apart (e.g., .22 and .39) than scores across two category boundaries (e.g., .39 and .42). These category boundaries were selected because of historical context, and are therefore meaningful in context. The reader should also note that the item difficulty and item discrimination indices are sample-specific. Since the samples in this study are not samples of applicants, but rather, officer cadets and enlisted personnel, the results may not be identical to those for operational samples.

As can be seen in Table 4, most items in Form 0 on this sample range in difficulty from .21 to .80. Four subtests have no item difficulties below .41 (Reading Comprehension, Math Knowledge, Block Counting, and Hidden Figures), while one subtest (Table Reading) has the majority of its item difficulties in the .81 to .99 category (25 of 40). Only two subtests have

Table 4. Distribution of Item Difficulty

Subtest	O[a]					P_1					P_2				
	.00-.20	.21-.40	.41-.60	.61-.80	.81-.99	.00-.20	.21-.40	.41-.60	.61-.80	.81-.99	.00-.20	.21-.40	.41-.60	.61-.80	.81-.99
Verbal Analogies	0	5	5	5	7	0	1	10	9	5	0	1	7	9	8
Arithmetic Reasoning	0	6	6	7	2	0	3	11	7	4	0	3	9	9	4
Reading Comprehension	0	0	11	13	1	0	5	8	11	1	0	1	10	11	3
Data Interpretation	1	4	10	6	2	0	1	13	8	3	0	1	6	15	3
Word Knowledge	0	3	9	10	2	0	2	9	11	3	0	2	7	14	2
Math Knowledge	0	0	11	13	1	0	0	9	13	3	0	0	14	9	2
Mechanical Comprehension	0	3	11	5	0	0	7	10	3	0	0	5	9	6	0
Electrical Maze[b]	0	4	9	7	0	0	10	5	5	0	0	10	3	7	0
Scale Reading[b]	0	2	22	13	2	0	1	13	20	6	0	2	17	17	4
Instrument Comprehension[b]	0	1	9	10	0	0	2	5	13	0	0	1	5	12	2
Block Counting[b]	0	0	5	10	5	0	0	4	10	6	0	2	3	9	6
Table Reading[b]	0	2	6	7	25	0	2	6	5	27	0	0	6	5	29
Aviation Information	0	7	10	3	0	0	8	9	3	0	0	7	7	5	1
Rotated Blocks	0	5	5	3	2	0	5	6	1	3	0	3	7	4	1
General Science	1[c]	5	12	1	1	0	5	10	4	1	0	5	11	3	1
Hidden Figures	0	0	6	3	6	0	0	6	5	4	0	2	5	2	6

[a] 12 items deleted from Form O scoring.
[b] Analyzed as a speeded test.
[c] $\underline{p}$ = .095.

items below .21, with each having just one item in that range. This indicates that the AFOQT has very few extremely difficult items and only one subtest (Table Reading) contains a majority of extremely easy items. As shown in Table 5, the median level of item difficulty for nine subtests falls in the .41 to .60 category, whereas six subtests have a median difficulty in the .61 to .80 category. Thus, most subtests are of average difficulty, with about one-third of the subtests being above average on the difficulty index (i.e., easier subtests).

The items in Form P_1 also fall mainly in the difficulty range of .21 to .80. A notable exception is Table Reading, with the majority of item difficulty values between .81 and .99 (27 of 40). This reveals that Table Reading is relatively easy. No subtest has items with difficulties below .21. Eight subtests in Form P_1 have medians in the .41 to .60 category and seven subtest difficulty medians fall between .61 and .80. This shows that about half of the subtests are of average difficulty and about half of the subtests are above average on the difficulty index (i.e., easier subtests).

The items in P_2 also fall mainly in the range from .21 to .80. As with Form P_1, most of the items in Table Reading fall in the .81 to .99 category (29 of 40). As before, Table Reading appears to be one of the easier subtests. Only five subtests have medians in the .41 to .60 range, while nine fall in the .61 to .80 range. This indicates that about one-third of the subtests in P_2 are of average difficulty and about one-half are above average on the difficulty index.

The first set of comparisons among the three test forms focuses on changes in the distributions of item difficulty from Form 0 to Forms P_1 and P_2. Collectively, the distributional statistics in Tables 4 and 5 indicate that, relative to the items in Form 0, items in Forms P_1 and P_2 have shifted toward the easier end of the difficulty continuum. Four subtests in both Forms P_1 and P_2 consistently have higher mean item difficulty values (greater than .02 points), and usually higher median, minimum, and maximum values, than the same subtests in Forms 0. These subtests are Arithmetic Reasoning, Data Interpretation, Scale Reading, and General Science. An additional four subtests are easier in only one of the new forms: Block Counting in Form P_1 and Verbal Analogies, Word Knowledge, and Instrument Comprehension in Form P_2. Several exceptions to the trend toward easier items in Forms P are noteworthy. The difficulty of the items in the Aviation Information, Math Knowledge, and Table Reading subtests is comparable across forms. Further, two Form 0 subtests contain items which are easier on the average than those in either Form P_1 or P_2 (Electrical Maze and Hidden Figures). Three additional subtests in Form 0 are easier than those in Form P_1 only (Reading Comprehension, Mechanical Comprehension, and Rotated Blocks).

The second set of comparisons addresses the comparability of difficulty of the items in the new test forms. As shown in Tables 4 and 5, the difficulty level of items in about one-third of the subtests is highly similar in Forms P_1 and P_2 (Arithmetic Reasoning, Electrical Maze, Scale Reading, Table Reading, General Science, and Hidden Figures). In eight of the 10 remaining subtests, Form P_2 clearly contains more items of lower difficulty. The trend toward easier items in Form P_2 is most pronounced in the Reading Comprehension, Data Interpretation, Mechanical Comprehension, Instrument Comprehension, and Aviation Information subtests. The same pattern is evident but the difference in average item difficulty between Forms P_1 and P_2 is less in the Verbal Analogies, Word Knowledge, and Rotated Blocks subtests. Results indicate that only two subtests are easier in Form P_2 than in Form P_1 (Math Knowledge and Block Counting). Test equating procedures were applied later to ensure that equivalent scale scores were derived for Forms P_1 and P_2, despite the observed differences in item difficulty.

Item Discrimination. Item discrimination was operationally defined as the biserial correlation between the score on an individual item (0 = incorrect, 1 = correct) and the subtest

Table 5. Summary Statistics of Item Difficulty

Subtest	O[a]					P_1					P_2				
	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD
Verbal Analogies	.66	.73	.33	.94	.21	.66	.69	.37	.97	.18	.69	.69	.38	.97	.17
Arithmetic Reasoning	.56	.58	.31	.82	.17	.61	.60	.34	.87	.16	.61	.61	.24	.90	.17
Reading Comprehension	.65	.67	.46	.82	.10	.59	.60	.28	.82	.17	.66	.67	.36	.89	.14
Data Interpretation	.51	.49	.15	.88	.17	.61	.57	.34	.88	.14	.65	.63	.35	.94	.13
Word Knowledge	.59	.60	.23	.87	.17	.61	.61	.34	.85	.15	.63	.65	.34	.88	.15
Math Knowledge	.64	.62	.48	.82	.10	.66	.64	.54	.87	.10	.63	.60	.44	.84	.11
Mechanical Comprehension	.52	.48	.29	.79	.14	.47	.43	.25	.73	.12	.51	.50	.28	.73	.13
Electrical Maze[b]	.53	.52	.27	.77	.13	.45	.42	.25	.70	.16	.47	.40	.25	.75	.18
Scale Reading[b]	.60	.57	.38	.89	.13	.64	.65	.32	.91	.15	.63	.64	.28	.89	.15
Instrument Comprehension[b]	.62	.61	.38	.73	.10	.62	.63	.33	.76	.11	.66	.68	.37	.84	.12
Block Counting[b]	.68	.72	.42	.91	.15	.73	.74	.52	.92	.12	.69	.74	.26	.92	.18
Table Reading[b]	.78	.86	.33	.93	.16	.78	.87	.39	.93	.17	.80	.86	.43	.94	.15
Aviation Information	.48	.48	.28	.75	.14	.46	.46	.24	.81	.15	.50	.50	.29	.90	.17
Rotated Blocks	.55	.51	.30	.87	.19	.51	.49	.22	.86	.22	.54	.51	.29	.86	.18
General Science	.46	.47	.10	.86	.16	.51	.49	.30	.86	.14	.51	.50	.29	.86	.15
Hidden Figures	.70	.70	.41	.94	.20	.66	.64	.41	.94	.19	.66	.63	.35	.91	.21

[a] 12 items deleted from Form O scoring.
[b] Analyzed as a speeded test.

total score. Items with discrimination values below .21 are typically viewed as having poor discriminative power while items above .81 are viewed as having excellent discriminative power. Item discrimination data are presented in Tables 6 and 7.

The majority of the item discrimination values for Form 0 fall in the .41 to .80 range, suggesting that most items have average to above average discriminative power. Verbal Analogies, Reading Comprehension, Math Knowledge, and Table Reading each have several items with discrimination values between .81 and .99. While no subtest has items with values below .21, seven subtests have at least one item in the .21 to .40 range (Scale Reading has 8 of 39 items in the latter category). The median discrimination value for twelve subtests falls in the .61 to .80 category, indicating the items as a whole have good discrimination abilities. Four subtest median discrimination values (Data Interpretation, Mechanical Comprehension, Scale Reading, and General Science) are between .41 and .60, indicating moderate discrimination abilities.

The discrimination pattern for Form P_1 is somewhat similar to that for Form 0, with the majority of values falling in the .41 to .80 range. While six subtests have at least one highly discriminating item (i.e., in the .81 to .99 range), nine have items of below average discriminative power (i.e., items in the .21 to .40 category). Scale Reading has 10 of 40 items in the latter category. Five subtests (Verbal Analogies, Data Interpretation, Mechanical Comprehension, Electrical Maze, and Scale Reading) have median values in the .41 to .60 range while eleven subtests have medians in the .61 to .80 range. On the whole, Form P_1 items have a good ability to make discriminations among individuals.

Most Form P_2 items have discrimination values in the .61 to .80 category. Eight subtests have at least one item in the .81 to .99 range, with Math Knowledge and Instrument Comprehension having approximately 50% of the items in that range. Eight subtests have items in the below average category, with Scale Reading having 8 of its 40 items there. Only three subtests have median discrimination values in the .41 to .60 category (Mechanical Comprehension, Electrical Maze, and Scale Reading), whereas thirteen subtests have median discrimination values spread between .61 and .80. In short, Form P_2 also shows good discrimination ability.

In general, the distributions of item discrimination values for Forms P_1 and P_2 are either similar to Form 0 distributions or tend to shift to higher levels of discrimination. In seven subtests, the mean and median discrimination values for items in both Forms P_1 and P_2 exceed those for items in Form 0 by at least .03 (Arithmetic Reasoning, Data Interpretation, Word Knowledge, Math Knowledge, Mechanical Comprehension, Scale Reading, and General Science). The same result is seen for two additional subtests in Form P_2 only (Instrument Comprehension and Aviation Information). Items in the same subtests on the other new test, Form P_1, are comparable in discriminative power to that of items in Form 0. Of the remaining seven subtests, only Electrical Maze items are clearly superior in discriminability on Form 0. The rest of the subtests are either similar among the three forms (Hidden Figures) or provide somewhat better discrimination in Form 0 than in one (but not both) of the new forms (Verbal Analogies, Reading Comprehension, Block Counting, Table Reading, and Rotated Blocks).

A comparison between Forms P_1 and P_2 shows that most subtests are composed of items with highly similar discriminative power. The consistency is observed in the distribution of item discrimination values in Table 6 and in the summary statistics in Table 7. In ten subtests the mean discrimination values for the two new forms differ by .02 or less. In the other six subtests, one test form is clearly superior to the other. Both the mean and median item discrimination values are higher in the Block Counting and Table Reading subtests on Form P_1 and in the Reading Comprehension, Data Interpretation, Instrument Comprehension, and Rotated Blocks subtests on Form P_2. Although Forms P_1 and P_2 are not precisely equivalent in item discrimination, the purpose of the follow-on test equating analyses is to ensure that converted test scores will be directly equivalent.

Table 6. Distribution of Item Discrimination

Subtest	O[a]				P_1				P_2			
	.21-.40	.41-.60	.61-.80	.81-.99	.21-.40	.41-.60	.61-.80	.81-.99	.21-.40	.41-.60	.61-.80	.81-.99
Verbal Analogies	0	9	10	3	0	13	12	0	0	11	11	3
Arithmetic Reasoning	0	6	15	0	0	2	19	4	0	2	20	3
Reading Comprehension	0	4	16	5	1	5	19	0	0	5	17	3
Data Interpretation	3	16	4	0	3	12	10	0	1	9	14	1
Word Knowledge	1	5	16	2	1	1	21	2	0	2	19	4
Math Knowledge	0	5	14	6	0	3	12	10	0	4	9	12
Mechanical Comprehension	4	11	4	0	2	12	6	0	2	11	7	0
Electrical Maze[b]	1	6	13	0	2	13	5	0	1	11	8	0
Scale Reading[b]	8	28	3	0	10	16	14	0	8	25	7	0
Instrument Comprehension[b]	0	7	12	1	0	4	14	2	0	3	9	8
Block Counting[b]	0	6	14	0	1	6	13	0	2	6	12	0
Table Reading[b]	1	8	28	3	2	7	26	5	3	10	25	2
Aviation Information	0	10	10	0	0	10	10	0	1	8	11	0
Rotated Blocks	0	4	11	0	0	5	10	0	0	3	12	0
General Science	4	8	8	0	2	7	11	0	1	8	11	0
Hidden Figures	0	2	13	0	0	1	13	1	0	2	13	0

[a] 12 items deleted from Form O scoring.
[b] Analyzed as a speeded test.

Table 7. Summary Statistics of Item Discrimination

Subtest	O[a]					P_1					P_2				
	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD
Verbal Analogies	.65	.64	.47	.84	.12	.62	.60	.50	.80	.08	.64	.63	.42	.83	.12
Arithmetic Reasoning	.65	.66	.50	.75	.08	.71	.71	.45	.84	.09	.71	.71	.50	.86	.08
Reading Comprehension	.71	.72	.47	.87	.11	.65	.68	.31	.80	.12	.69	.71	.48	.84	.09
Data Interpretation	.52	.54	.30	.69	.10	.57	.59	.30	.79	.12	.63	.63	.26	.81	.12
Word Knowledge	.67	.68	.41	.83	.11	.70	.72	.40	.85	.09	.72	.72	.43	.90	.11
Math Knowledge	.72	.71	.49	.92	.11	.78	.80	.56	.96	.11	.77	.80	.57	.94	.11
Mechanical Comprehension	.51	.52	.27	.71	.12	.56	.56	.36	.71	.09	.58	.59	.35	.75	.11
Electrical Maze[b]	.62	.61	.35	.78	.10	.53	.54	.29	.67	.10	.55	.58	.29	.70	.11
Scale Reading[b]	.48	.47	.29	.71	.10	.52	.54	.23	.74	.13	.51	.52	.32	.74	.11
Instrument Comprehension[b]	.64	.65	.43	.83	.11	.66	.65	.49	.86	.10	.74	.76	.47	.92	.13
Block Counting[b]	.63	.66	.46	.78	.09	.64	.65	.39	.79	.12	.59	.63	.34	.76	.13
Table Reading[b]	.68	.71	.32	.84	.12	.69	.73	.38	.83	.13	.63	.67	.29	.85	.13
Aviation Information	.60	.61	.42	.75	.09	.61	.62	.46	.76	.09	.63	.64	.40	.75	.09
Rotated Blocks	.66	.66	.51	.77	.08	.63	.62	.52	.79	.07	.66	.69	.51	.77	.08
General Science	.53	.53	.30	.74	.15	.61	.63	.34	.76	.11	.59	.63	.31	.76	.12
Hidden Figures	.69	.71	.57	.81	.07	.70	.71	.54	.81	.07	.69	.71	.52	.79	.07

[a]12 items deleted from Form O scoring.
[b]Analyzed as a speeded test.

Subtest Analysis

The format for the discussion of the subtest analyses will be similar to the preceding format in that each form of the AFOQT will be discussed independently of the other forms. For each subtest, the discussion will focus on the proportion correct[4] and a measure of internal consistency. Skew, kurtosis, and the intercorrelations of the subtests will be discussed for all three forms together. These data are presented in Table 8 and Table 9.

As discussed earlier in this paper, items were omitted from scoring in six Form O subtests due to miskeys or poor item performance. To facilitate across-forms comparisons of the affected subtests, two mean scores are reported for Form O (see Table 8). The first is the actual mean number of items answered correctly and is based on the number of items scored. The second set of mean scores (shown in parentheses) is adjusted for subtest length. Ratios were solved to determine the Form O mean value for a subtest length equivalent to that of Forms P_1 and P_2.

The reader should also note that some of the reliability indices reported in Table 8 may be inflated because of the speeded nature of the subtests. The reliability values for Electrical Maze, Scale Reading, Instrument Comprehension, Block Counting, and Table Reading are not reported, because no parallel form indices are yet available. A more appropriate measure of reliability would involve the use of correlation between separately timed parallel forms. These data are not available at this time.

Form O has five subtests for which the average proportion of items answered correctly is greater than .60 and three subtests with average proportions less than .50. The remaining subtests fall between .50 and .60. This shows that most subtests are average to below average in difficulty. Hidden Figures and Table Reading are the two easiest subtests while Electrical Maze, General Science, and Aviation Information are the most difficult. The measures of internal consistency (reliability) are fairly high for Form O. One subtest is below .71, four subtests fall in the range of .71 to .80, and six fall in the range of .81 to .90. Five subtests are judged to have speeded properties (Electrical Maze, Instrument Comprehension, Scale Reading, Table Reading, and Block Counting); therefore, internal consistency is not an appropriate measure of reliability for these subtests.

Form P_1 has eight subtests with average proportion correct values greater than .60 and three subtests with values less than .50. The remaining five subtests have proportions between .50 and .60. This shows that most subtests are average to below average in difficulty (i.e., easy subtests). In Form P_1, Table Reading and Math Knowledge are the easiest subtests while Electrical Maze, Aviation Information, and Mechanical Comprehension are the most difficult. The measures of internal consistency are fairly high for Form P_1. The values of internal consistency measures fall mainly between .81 and .90 (seven subtests). Three subtests have reliability values between .71 and .80, and one subtest has a value greater than .91. As with Form O, five subtests are judged to be speeded; therefore, other measures of reliability need to be generated.

Form P_2 also has eight subtests with average proportion correct values greater than .60, but only one subtest with a value less than .50. The remaining seven subtests have proportions between .51 and .60. This shows that most subtests are average to below average in difficulty

---

[4] The proportion correct values reported in Table 8 are the same as the mean item difficulty values shown in Table 5 for power subtests but not for speeded subtests. Proportion correct values were computed by dividing the mean number of items answered correctly by the number of items scored.

Table 8. Descriptive Statistics of Subtests

Subtest	Proportion correct 0	Proportion correct P_1	Proportion correct P_2	Mean 0[a]	Mean P_1	Mean P_2	Standard deviation 0	Standard deviation P_1	Standard deviation P_2	Skew 0	Skew P_1	Skew P_2	Kurtosis 0	Kurtosis P_1	Kurtosis P_2	Reliability 0	Reliability P_1	Reliability P_2
Verbal Analogies	.657	.659	.691	14.46 (16.43)	16.48	17.27	4.32	4.85	4.90	-.65	-.52	-.70	-.23	-.46	-.29	.819	.829	.840
Arithmetic Reasoning	.558	.605	.615	11.72 (13.95)	15.13	15.37	4.85	6.24	6.16	-.06	-.23	-.24	-.99	-1.08	-1.02	.846	.899	.898
Reading Comprehension	.654	.586	.655	16.35	14.66	16.38	6.28	5.76	5.90	-.49	-.26	-.49	-.86	-.86	-.82	.899	.874	.888
Data Interpretation	.515	.611	.651	11.84 (12.87)	15.28	16.27	4.33	5.15	5.44	.07	-.31	-.45	-.62	-.78	-.74	.762	.826	.857
Word Knowledge	.588	.614	.625	14.12 (14.71)	15.35	15.62	5.74	6.19	6.27	-.13	-.24	-.35	-.96	-1.00	-.54	.890	.896	.901
Math Knowledge	.637	.662	.626	15.93	16.54	15.64	6.46	6.84	6.96	-.37	-.43	-.34	-1.06	-1.18	-1.23	.904	.923	.922
Mechanical Comprehension	.517	.468	.505	9.82 (10.34)	9.35	10.10	3.62	4.27	4.36	-.03	.23	.01	-.68	-.86	-.89	.703	.781	.794
Electrical Maze[b]	.465	.392	.413	9.30	7.84	8.25	4.71	3.88	3.91	.47	.48	.39	-.60	.02	-.29			
Scale Reading[b]	.557	.591	.580	21.74 (22.30)	23.62	23.20	7.34	7.73	7.65	-.22	-.24	-.22	-.64	-.79	-.65			
Instrument Comprehension[b]	.553	.544	.594	11.05	10.87	11.87	5.31	5.47	5.69	-.05	-.08	-.25	-1.16	-1.17	-1.26			
Block Counting[b]	.597	.639	.599	11.93	12.78	11.97	4.57	4.45	4.14	-.37	-.55	-.43	-.57	-.26	-.28			
Table Reading[b]	.678	.683	.696	27.10	27.33	27.84	8.14	7.93	7.76	-.63	-.59	-.53	.16	.12	-.10			
Aviation Information	.480	.462	.505	9.60	9.24	10.10	4.50	4.59	4.56	.39	.47	.28	-.72	-.69	-.77	.812	.825	.828
Rotated Blocks	.545	.513	.537	8.18	7.70	8.05	3.44	3.25	3.54	-.21	-.05	-.10	-.73	-.81	-.83	.783	.759	.793
General Science	.465	.509	.510	9.29	10.18	10.20	3.89	4.59	4.38	.24	.09	-.02	-.64	-.91	-.93	.740	.822	.801
Hidden Figures	.695	.661	.655	10.42	9.92	9.83	3.04	3.35	3.19	-.61	-.29	-.38	-.05	-.71	-.38	.765	.802	.780

[a]Values shown in parentheses are mean raw scores for Form 0 subtests adjusted to the item length of the same subtests in Forms P. The other values reported (not in parentheses) are actual mean scores based on number of items scored.

[b]Analyzed as a speeded subtest. Therefore internal consistency is not appropriate.

Table 9. Intercorrelations Among Subtests

Subtests		AR	RC	DI	WK	MK	MC	EM	SR	IC	BC	TR	AI	RB	GS	HF
VA	0	.62	.75	.58	.72	.67	.56	.38	.55	.49	.51	.44	.45	.48	.56	.50
	P_1	.67	.71	.65	.73	.64	.55	.38	.56	.49	.48	.44	.45	.47	.61	.47
	P_2	.69	.76	.70	.76	.67	.55	.40	.53	.52	.50	.51	.45	.51	.64	.50
AR	0		.63	.68	.51	.77	.55	.47	.71	.50	.56	.49	.37	.55	.54	.49
	P_1		.64	.77	.59	.77	.56	.42	.72	.46	.56	.54	.39	.50	.62	.48
	P_2		.65	.78	.58	.80	.61	.48	.68	.56	.56	.57	.42	.60	.65	.52
RC	0			.61	.80	.67	.54	.39	.52	.45	.47	.44	.46	.45	.62	.45
	P_1			.65	.75	.55	.46	.33	.54	.42	.45	.43	.42	.37	.58	.38
	P_2			.69	.78	.60	.52	.33	.49	.45	.45	.46	.43	.43	.62	.44
DI	0				.53	.65	.49	.44	.63	.48	.54	.49	.38	.45	.48	.46
	P_1				.57	.66	.51	.42	.70	.48	.55	.56	.42	.48	.56	.49
	P_2				.59	.71	.58	.43	.66	.52	.54	.58	.41	.57	.60	.51
WK	0					.57	.49	.31	.41	.42	.39	.36	.44	.35	.58	.38
	P_1					.55	.46	.30	.46	.38	.40	.34	.43	.36	.63	.35
	P_2					.57	.52	.29	.42	.46	.41	.40	.44	.41	.64	.42
MK	0						.55	.48	.68	.48	.58	.54	.38	.57	.59	.53
	P_1						.54	.41	.65	.42	.52	.57	.35	.51	.66	.49
	P_2						.56	.46	.63	.55	.52	.56	.40	.56	.66	.55
MC	0							.50	.50	.56	.51	.35	.54	.57	.62	.44
	P_1							.49	.50	.56	.48	.32	.57	.58	.66	.40
	P_2							.52	.48	.62	.46	.38	.59	.60	.72	.45
EM	0								.50	.51	.56	.41	.38	.52	.46	.44
	P_1								.50	.49	.50	.38	.36	.44	.45	.42
	P_2								.52	.51	.52	.43	.39	.45	.45	.45
SR	0									.55	.63	.58	.40	.55	.47	.55
	P_1									.50	.64	.61	.37	.48	.51	.48
	P_2									.54	.64	.65	.37	.51	.47	.50
IC	0										.58	.44	.64	.55	.52	.51
	P_1										.52	.39	.62	.51	.52	.41
	P_2										.54	.48	.64	.58	.60	.52
BC	0											.60	.40	.61	.45	.59
	P_1											.56	.38	.56	.47	.46
	P_2											.62	.37	.55	.44	.52
TR	0												.32	.45	.37	.48
	P_1												.27	.39	.39	.43
	P_2												.37	.43	.39	.48
AI	0													.41	.54	.36
	P_1													.40	.53	.28
	P_2													.43	.57	.40
RB	0														.51	.52
	P_1														.52	.49
	P_2														.57	.55
GS	0															.41
	P_1															.41
	P_2															.46

(i.e., easy subtests). For Form P_2, Table Reading and Verbal Analogies are the easiest subtests while Electrical Maze, Aviation Information, and Mechanical Comprehension are the most difficult. As for the measures of internal consistency, Form P_2 has four subtests with values in the .71 to .80 range, six subtests with values in the .81 to .90 range, and one subtest with a value greater than .91. Again, those tests judged to be speeded do not have meaningful values at this time.

For the comparisons between forms, the subtests were judged to be similar in difficulty if the actual (or adjusted) mean scores differed by less than one raw score unit[5]. Three subtests in Form 0 were more difficult than the corresponding subtests in Form P_1 (Arithmetic Reasoning, Data Interpretation, and Scale Reading) while two subtests were easier in Form 0 than in Form P_1 (Reading Comprehension and Electrical Maze). The remaining 11 subtests were judged to be similar in mean scores. Form 0 had two subtests that were more difficult than the corresponding subtests in Form P_2 (Arithmetic Reasoning and Data Interpretation) and only one subtest that was easier (Electrical Maze). As for the comparisons of internal consistency measures, Forms P_1 and P_2 were very similar to Form 0. For all three forms, most reliability values were in the .81 to .90 range, with Math Knowledge having the highest internal consistency values. Forms P_1 and P_2 had slightly higher values than Form 0. This indicates that the new forms may be slightly more internally consistent than the previous form.

Form P_2 is slightly easier than Form P_1. Two subtests in P_2 were easier than their counterparts in P_1 (Reading Comprehension and Instrument Comprehension); none of the subtests in P_2 was more difficult than its counterpart in P_1. The remaining 14 subtests had mean scores that differed by less than one raw score point. The test equating analyses to be described later in this paper have the effect of removing observed differences in subtest difficulty from test scores. Thus, the test form administered becomes a matter of indifference to examinees. As for internal consistency, inspection of Table 8 reveals that the two forms are almost identical.

The discussion of skew and kurtosis for the subtests was delayed until this point, because the pattern of results is nearly identical for the three forms. It should be pointed out here that the normal distribution has a value of 0.0 for both skew and kurtosis. For all three forms, no subtest exhibited skew values less than -1.00 or greater than 1.00. This indicates that the subtests are relatively symmetrical. As for kurtosis, twelve subtests tend toward normality. Four subtests have kurtosis values around -1.00 or less. These subtests (Arithmetic Reasoning, Word Knowledge, Math Knowledge, and Instrument Comprehension) have slightly flatter distributions than do the remaining subtests.

The intercorrelations for Forms 0 and P are presented in Table 9. Rather than presenting a separate table for each of the three correlation matrices, the data are presented in one table for easier comparison across forms. Please note that the tabled values are for correlations between two subtests for one form; they are not correlations between forms (e.g., Form 0 with Form P_1). The highest correlations are between Arithmetic Reasoning and Math Knowledge and between Reading Comprehension and Word Knowledge. The former pair of subtests are in the Quantitative composite and the latter are in the Verbal composite. The lowest correlations are found between Electrical Maze and Reading Comprehension, Electrical Maze and Word Knowledge, Table Reading and Mechanical Comprehension, and Table Reading and Aviation Information. There is a great amount of consistency in correlations across the three forms. Most of the differences

[5]This difference was chosen as a standard because one raw score point can have operational implications. For example, in some portions of the Verbal composite conversion table, a difference of one raw score unit results in a difference of three percentile units.

among the triads of correlations fall within expected ranges, given the reliabilities of the subtests. The largest difference between any corresponding pair of correlations is .13. This occurs for the correlations between Math Knowledge and Instrument Comprehension for Forms P_1 and P_2 and the correlations between Block Counting and Hidden Figures for Forms P_1 and 0. Nonetheless, there is a high degree of similarity in correlation matrices across the three forms.

Composite Analysis

The format for the discussion of the composite analyses differs from the preceding format in that the three forms of the AFOQT are not discussed individually. The comparison of Form 0 to Forms P_1 and P_2 is followed by the comparison of Forms P_1 and P_2. For each composite, the discussion will focus on the average proportion of items answered correctly, actual composite mean scores, and composite mean scores adjusted for item length. The skew and kurtosis of the composite distributions and the composite intercorrelations are also discussed.

As a result of the 12 items being removed from the scoring of Form 0, Forms P_1 and P_2 and Form 0 have a different number of items contributing to the composite scores. Table 10 compares the number of items for Form 0 and Forms P_1 and P_2 by composite.

Table 10. Number of Items in AFOQT Forms 0 and P Composites

Composite	Test form: 0	P_1 & P_2
Navigator-Technical	257	265
Pilot	200	205
Academic Aptitude	140	150
Verbal	71	75
Quantitative	69	75

Table 11 presents descriptive data at the composite level on which the following discussion is based. The proportion correct values for Form 0 composites tend to be lower than those for both Forms P_1 and P_2. This generalization holds for the Navigator-Technical, Academic Aptitude, and Quantitative composites. For these composites the differences between Form 0 adjusted mean scores and Form P_1 and P_2 actual mean scores exceed one point. The generalization also holds for the comparison of the Pilot and Verbal composites for Forms 0 and P_2. It does not hold, however, for two cases in the comparison of Forms 0 and P_1; the differences in proportion correct values for the Pilot and Verbal composites do not translate to mean score differences in excess of one point. It should be noted here that the effects of the differences in test difficulty are removed by the equating process.

The proportion correct values for Forms P_1 and P_2 are most similar for the Navigator-Technical and Quantitative composites. However, an inspection of the mean scores indicates that the only composite on which Forms P_1 and P_2 differ by less than one raw score point is the Quantitative composite. The order of the other composites in terms of magnitude of mean score differences (least to greatest) is Navigator-Technical, Verbal, Pilot, and Academic Aptitude. A particularly noteworthy finding is that Form P_2 has a higher mean score on all composites than does Form P_1, indicating that Form P_2 is the easier of the new AFOQT forms. However, after equating there should be no significant difference in scores between Forms P_1 and P_2.

Table 11. Composite Descriptive Statistics

AFOQT composite	Proportion			Mean			Standard deviation			Skew			Kurtosis		
	0	P_1	P_2	0[a]	P_1	P_2	0	P_1	P_2	0	P_1	P_2	0	P_1	P_2
Navigator-Technical	.573	.587	.592	147.27 (151.85)	155.63	156.73	41.56	44.23	44.79	-.34	-.35	-.39	-.69	-.78	-.70
Pilot	.575	.573	.588	114.99 (117.86)	117.48	120.61	31.87	32.01	32.61	-.39	-.37	-.41	-.63	-.66	-.59
Academic Aptitude	.603	.622	.643	84.43 (90.46)	93.41	96.56	27.01	29.66	30.61	-.44	-.43	-.50	-.72	-.82	-.73
Verbal	.633	.620	.657	44.93 (47.46)	46.48	49.27	14.99	15.25	15.71	-.46	-.36	-.52	-.70	-.76	-.71
Quantitative	.572	.626	.630	39.50 (42.93)	46.93	47.28	14.05	16.58	17.07	-.20	-.33	-.33	-.98	-1.06	-1.03

[a]Values shown in parentheses are mean raw scores for Form O composites adjusted to the item length of the Forms P composites. The other values reported in the column (not in parentheses) are actual mean scores based on the number of items scored.

The skew and kurtosis for the composite distributions are highly similar for the three forms. For all three forms, all composites exhibited skew values between -.52 and -.20, indicating that the composite score distributions are relatively symmetrical. As for kurtosis, four composites exhibited values between -.59 and -.82, indicating that the distributions tend toward normality. The Quantitative composite had values of approximately -1.00, indicating slightly flatter distributions than those for the other composites.

Correlations among the five composites are shown in Table 12. The correlations are highly similar for the three forms. That is, the intercorrelation matrix for Form O is similar to those of Forms P_1 and P_2, with the latter two being nearly identical. The lowest correlations were found between the Verbal composite and each of the Pilot, Navigator-Technical, and Quantitative composites. It should be noted that the correlations between the Academic Aptitude composite and both the Verbal and Quantitative composites are artificially inflated, because the Academic Aptitude composite is a linear combination of the Verbal and Quantitative composites. The high correlation between the Pilot and Navigator-Technical composites is also inflated, because the composites have several subtests in common.

Table 12. **Intercorrelations Among Composites**

		Navigator-Technical	Academic Aptitude	Verbal	Quantitative
	O	.96	.80	.70	.80
Pilot	P_1	.95	.82	.71	.81
	P_2	.95	.82	.71	.81
	O		.87	.71	.91
Navigator-	P_1		.89	.72	.93
Technical	P_2		.89	.73	.93
	O			.94	.93
Academic	P_1			.93	.94
Aptitude	P_2			.93	.94
	O				.73
Verbal	P_1				.74
	P_2				.74

Note. The correlations are inflated, since the composites have several subtests in common.

Equating

Form N currently serves as the form on which the normative sample was constructed for the AFOQT. Forms P_1 and P_2 were equated to Form O in this study because Form O had been equated to Form N in previous research (see Rogers, Roach, & Wegner, 1986). This places scores for Forms P_1 and P_2 on the Form N metric.

As discussed earlier, decisions were made as to which of the four equating methods calculated was most appropriate for each of the five composites of Forms P_1 and P_2. Again, selection was based on the similarity of the distributions for the new test composite and the corresponding composite in the reference test, Form O. The Standard Error of Estimate (SEE) was used as a

goodness-of-fit measure to determine which smoothing method would be chosen for each of the equipercentile equatings. If one method resulted in a significantly smaller SEE than another, the method with the smaller SEE was chosen. When two forms of smoothing did not differ greatly, the form with the least complex regression equation was chosen. Table 13 contains the SEE values for the three forms of smoothing of the equipercentile equatings.

Table 13. Standard Error of Estimate for Linear, Quadratic, and Cubic Smoothing of Equipercentile Equating

AFOQT composite	P_1 equipercentile			P_2 equipercentile		
	Linear	Quadratic	Cubic	Linear	Quadratic	Cubic
Pilot	2.79	1.41	1.24	2.41	1.15	1.10
Navigator-Technical	3.15	1.85	1.82	3.47	1.89	1.88
Academic Aptitude	2.24	1.76	1.50	2.72	1.78	1.65
Verbal	1.42	1.40	.55	1.05	.64	.44
Quantitative	1.49	.78	.56	1.60	.79	.76

The equipercentile equating method was chosen for all equatings.[6] Further, the selection of polynomial smoothing method for each composite was governed by its joint performance on both of the new test forms. Thus, the type of smoothing might vary among the composites but not on any single composite for both Forms P_1 and P_2. For the Pilot, Navigator-Technical, and Quantitative composites, a quadratic polynomial smoothing method was selected. For these composites, the SEE values decreased significantly from linear to quadratic forms of smoothing, but only trivially to cubic. For the Academic Aptitude and Verbal composites, sufficient decreases in SEE were found from the quadratic to the cubic polynomial smoothing for both forms. Although the decrease for Form P_2 is less than the decrease for Form P_1, the cubic smoothing method was selected for both composites for the sake of logical consistency.[7]

Upon further inspection of the equatings, it was determined that separate conversion tables were needed for Forms P_1 and P_2. The equivalent raw scores on Forms P_1 and P_2 equated to different raw scores on Form O due to the differences between Forms P_1 and P_2. Although the forms were developed to be parallel in content, format, and so on, slight differences in raw score distributions were apparent. For example, in the Pilot composite in the range of the 10th through 90th percentiles, the raw scores for the same percentile on the two forms differed from 2

[6]Equipercentile equatings were selected in lieu of linear (z-score) equatings. For all composites on both Forms P_1 and P_2, the indexes of deviation (bias, average absolute deviation, and root mean square deviation) between the accepted polynomial smoothing and linear (z-score) equating were usually greater than .5 raw score points and often as large as 2 to 3 points. The magnitudes of these differences were judged too large to allow linear equatings of the tests.

[7]All equatings will be reviewed and evaluated in the IOT&E and new tables provided as dictated by the data.

to 4 points. Therefore, separate tables were required. The tables in Appendix A convert raw scores for each composite of each version of Form P to the percentile score based on the Form N metric.

IV. CONCLUSIONS AND RECOMMENDATIONS

The goals of this research were to determine the adequacy of items in Forms P using a more representative sample, to determine if Forms O, P_1, and P_2 are parallel, and to derive scores on Forms P that are comparable to scores on Form O.

Enlisted personnel (BMTS subjects) and prospective officers (cadets in AFROTC and OTS) were used to collect data for the purpose of examining item performance in Forms P. Based on the item difficulty results, it can be concluded that the items in Forms P are acceptable, since the majority of items fall within a desirable range of difficulty. Even though Forms P are slightly easier than Form O, the item difficulty distributions are similar enough to proceed with test equating in order to remove the effects of small differences in difficulty.

Based on item discrimination results, the items in Forms P are acceptable. On the whole, items in Forms P have similar or slightly higher discrimination values than those of Form O. The similarity across forms is even greater when comparing Form P_1 with Form P_2. It can be concluded that the new forms have a slightly better ability to discriminate among examinees of differing ability levels.

The majority of subtests have similar mean scores across the three forms. Form O has more difficult subtests in three cases and easier subtests in two cases, but the three forms provide almost identically shaped score distributions. The skew and kurtosis values indicated that the majority of subtests are symmetrical and tend toward normality. Furthermore, there is great consistency in these values across the three forms. The three forms of the AFOQT also show great consistency in the intercorrelation matrices. Therefore, it can be concluded that at the subtest raw score level, the three forms are generally parallel.

At the composite raw score level, the forms are also generally parallel. The proportion correct values for the composites tend to be higher for Forms P than Form O. This holds for all five composites of Form P_2 and for three composites of Form P_1. The proportion correct values for the remaining two composites of Form P_1 are similar. All composite score distributions are roughly symmetric, with moderate peaks. These three forms of the AFOQT are moderately parallel and, therefore, appropriate for equating.

Given that Forms P are generally parallel to Form O, it was possible to derive scores on Forms P which are equivalent to scores on Form O. Equipercentile equatings with either quadratic or cubic smoothings were used to generate provisional conversion tables (Appendix A). For three of the composites (Pilot, Navigator-Technical, and Quantitative), quadratic smoothing was selected; for the remaining two composites (Verbal and Academic Aptitude), cubic smoothing was selected. The method of smoothing selected for each composite was the same for Forms P_1 and P_2. For example, quadratic smoothing was selected for the Pilot composites of both Forms P_1 and P_2. These tables are recommended for use operationally until the completion of an Initial Operational Test and Evaluation (IOT&E).

The goal of the planned IOT&E is to verify the conversion tables generated by this pre-implementation research. The methodology will be similar to that described in this paper in that Forms O and P will be distributed alternately within each testing session at each testing site. The data editing and analysis will resemble those reported here but will produce conversion tables based on operational data. The extent of the changes, if any, that will need to be made to the final Forms P operational conversion tables cannot be determined at this time.

REFERENCES

Air Force Regulation 35-8. (1983). Air Force military personnel testing system. Washington, DC: Department of the Air Force.

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, DC: American Council on Education.

Braun, H.I., & Holland, P.W. (1982). Observed score test equating: A mathematical analysis of some ETS equating procedures (pp. 9-49). In P.W. Holland & D.B. Rubin (Eds.), Test equating. New York: Academic Press.

Gould, R. B. (1978). Air Force Officer Qualifying Test Form N: Development and standardization (AFHRL-TR-78-43, AD-A059 746). Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.

Gulliksen, H. (1950). Theory of mental tests. New York: John Wiley & Sons, Inc.

Henrysson, S. (1971). Gathering, analyzing and using data on test items. In R. L. Thorndike (Ed.), Educational measurement (pp. 130-159). Washington, DC: American Council on Education.

Miller, R. E. (1974). Development and standardization of the Air Force Officer Qualifying Test Form M (AFHRL-TR-74-16, AD-778 837). Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.

Rogers, D.L., Roach, B.W., & Short, L.O. (1986). Mental ability testing in the selection of Air Force officers: A brief historical overview (AHFRL-TP-86-23, AD-A173 484). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Rogers, D.L., Roach, B.W., & Wegner, T.G. (1986). Air Force Officer Qualifying Test Form O: Development and standardization (AFHRL-TR-86-24, AD-A172 037). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Skinner, J., & Ree, M.J. (1987). Air Force Officer Qualifying Test (AFOQT): Item and factor analysis of Form O (AFHRL-TR-86-68, AD-A184 975). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Wegner, T.G., & Ree, M.J. (1985). Armed Services Vocational Aptitude Battery: Correcting the speeded subtests for the 1980 youth population (AFHRL-TR-85-14, AD-A158 823). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

APPENDIX A: PROVISIONAL CONVERSION TABLES FOR AFOQT P_1 AND P_2

Table A-1. AFOQT - P_1 Provisional Conversion Table for Pilot Composite

Raw	Percentile	Raw	Percentile
00 - 41	01	117	48
42 - 47	02	118	50
48 - 54	03	119	51
55 - 58	04	120	52
59 - 61	05	121	53
62 - 64	06	122	54
65 - 67	07	123	55
68 - 70	08	124	57
71	09	125	58
72 - 73	10	126	60
74 - 75	11	127	61
76 - 77	12	128	62
78 - 79	13	129	63
80	14	130	64
81	15	131	65
82	16	132	66
83 - 85	17	133	67
86	18	134	69
87	19	135	70
88 - 89	20	136	71
90	21	137	73
91	22	138	74
92	23	139	76
93 - 94	24	140	77
95	25	141	78
96	26	142	79
97	27	143	80
98	28	144	81
99	29	145	82
100	30	146	83
101	31	147 - 148	84
102	32	149 - 150	86
103	33	151	87
104	34	152	88
105	35	153	89
106	36	154	90
107	37	155	91
108	38	156	92
109	39	157	93
110	41	158	94
111	42	159 - 160	95
112	43	161 - 163	96
113	44	164 - 167	97
114	45	168 - 172	98
115	46	173 - 205	99
116	47		

Table A-2. AFOQT - P_1 Provisional Conversion Table for Navigator-Technical Composite

Raw	Percentile	Raw	Percentile
01 - 61	01	157	50
62 - 71	02	158	51
72 - 77	03	159	52
78 - 82	04	160	53
83 - 86	05	161	54
87 - 88	06	162	55
89 - 90	07	163	56
91 - 94	08	164	57
95 - 97	09	165	58
98 - 99	10	166	59
100 - 102	11	167	60
103 - 104	12	168	61
105 - 106	13	169	62
107 - 108	14	170	63
109 - 110	15	171	64
111 - 113	16	172 - 173	65
114 - 115	17	174	66
116 - 117	18	175	67
118	19	176	68
119 - 120	20	177	69
121 - 122	21	178	70
123 - 124	22	179	71
125	23	180	72
126	24	181 - 182	73
127 - 128	25	183	74
129	26	184	75
130	27	185	76
131	28	186	77
132	29	187	78
133 - 134	30	188 - 189	79
135	31	190	80
136	32	191 - 192	81
137	33	193	82
138	34	194 - 195	83
139	35	196	85
140 - 141	36	197	86
142	37	198 - 199	87
143 - 144	38	200 - 201	88
145	39	202 - 203	89
146	40	204 - 205	90
147	41	206 - 207	91
148	42	208	92
149	43	209 - 210	93
150	43	211 - 212	94
151	44	213 - 215	95
152	45	216 - 219	96
153	46	220 - 224	97
154	47	225 - 227	98
155	48	228 - 265	99
156	49		

Table A-3. AFOQT - P_1 Provisional Conversion Table for Academic Aptitude Composite

Raw	Percentile	Raw	Percentile
01 - 27	01	95	50
28 - 34	02	96	51
35 - 40	03	97	52
41 - 42	04	98	53
43 - 46	05	99	54
47 - 48	06	100	57
49 - 51	07	101	59
52	08	102	61
53 - 55	09	103	62
56 - 58	10	104	63
59	11	105	65
60	12	106	67
61	13	107	68
62	14	108	69
63	15	109	70
64 - 65	16	110	71
66	17	111	72
67 - 68	18	112	75
69 - 70	19	113	76
71	20	114	78
72	21	115	79
73	22	116	80
74	23	117	81
75	24	118	82
76	25	119	83
77	26	120	84
78	27	121	85
79	28	122	86
80	29	123	87
81	31	124	88
82	33	125 - 126	89
83	34	127	90
84	35	128	91
85	36	129	92
86	37	130 - 131	93
87	38	132	94
88	40	133 - 134	95
89	41	135 - 136	96
90	43	137 - 139	97
91	44	140 - 141	98
92	45	142 - 150	99
93	47		
94	49		

Table A-4. AFOQT - P_1 Provisional Conversion Table for Verbal Composite

Raw	Percentile	Raw	Percentile
01 - 13	01	43	41
14	02	44	46
15 - 17	03	45	48
18	04	46	50
19	05	47	53
20	06	48	55
21	07	49	57
22	08	50	60
23	09	51	62
24	10	52	64
25	11	53	67
26	12	54	72
27	13	55	74
28	14	56	77
29	15	57	78
30	17	58 - 59	81
31	18	60	84
32	19	61	86
33	21	62	87
34	23	63	90
35	24	64	92
36	26	65 - 66	93
37	30	67	96
38	32	68	97
39	33	69 - 70	98
40	36	71 - 75	99
41	38		
42	40		

Table A-5. AFOQT - P_7 Provisional Conversion Table for Quantitative Composite

Raw	Percentile	Raw	Percentile
01 - 14	01	49	52
15 - 17	02	50	54
18 - 20	03	51 - 52	57
21	04	53	59
22 - 23	05	54	61
24	06	55	64
25	08	56	66
26 - 27	09	57	69
28	10	58	71
29	11	59	75
30 - 31	14	60	76
32	15	61	78
33	17	62	80
34	19	63	82
35 - 36	21	64	85
37	24	65	86
38	26	66	88
39	28	67	90
40	31	68	92
41 - 42	33	69	93
43	34	70	94
44	38	71	95
45	41	72	96
46	43	73	97
47	45	74	98
48	48	75	99

Table A-6. AFOQT - P_2 Provisional Conversion Table for Pilot Composite

Raw	Percentile	Raw	Percentile
00 - 44	01	121	50
45 - 49	02	122	51
50 - 56	03	123	52
57 - 61	04	124	53
62 - 63	05	125	54
64 - 67	06	126	55
68 - 70	07	127	56
71 - 72	08	128	57
73	09	129	58
74 - 75	10	130	60
76 - 78	11	131	62
79 - 80	12	132	63
81 - 82	13	133	64
83	14	134	65
84	15	135	66
85	16	136	67
86 - 87	17	137	69
88	18	138	70
89	19	139	71
90 - 91	20	140	73
92	21	141	74
93 - 94	22	142	75
95	23	143	76
96 - 97	24	144	77
98	25	145	78
99	26	146	79
100	27	147	81
101	28	148	82
102	29	149	83
103	30	150 - 151	84
104	31	152	85
105	32	153 - 154	86
106	33	155	87
107	34	156	88
108	35	157	89
109	36	158	91
110	37	159	92
111	38	160	93
112	39	161 - 162	94
113	41	163 - 164	95
114	42	165 - 166	96
115	43	167 - 171	97
116	44	172 - 176	98
117	45	177 - 205	99
118	46		
119	47		
120	48		

Table A-7. AFOQT - P_2 Provisional Conversion Table for Navigator-Technical Composite

Raw	Percentile	Raw	Percentile
00 - 59	01	159	50
60 - 71	02	160	51
72 - 77	03	161	52
78 - 82	04	162	53
83 - 85	05	163	54
86 - 88	06	164	55
89 - 90	07	165	56
91 - 94	08	166	57
95 - 97	09	167	58
98 - 100	10	168	59
101 - 102	11	169	60
103 - 104	12	170	61
105 - 106	13	171	62
107 - 109	14	172	63
110 - 111	15	173	64
112 - 113	16	174 - 175	65
114 - 116	17	176	66
117 - 118	18	177	67
119	19	178	68
120 - 121	20	179	69
122 - 123	21	180	70
124 - 125	22	181	71
126	23	182	72
127	24	183 - 184	73
128 - 129	25	185	74
130	26	186	75
131	27	187	76
132	28	188	77
133	29	189	78
134 - 135	30	190 - 191	79
136 - 137	31	192	80
138	32	193 - 194	81
139	33	195	82
140	34	196	83
141	35	197	84
142	36	198	85
143	37	199	86
144 - 145	38	200 - 201	87
146	39	202 - 203	88
147	40	204 - 205	89
148	41	206 - 207	90
149	42	208 - 209	91
150 - 151	43	210	92
152	44	211 - 212	93
153 - 154	45	213 - 214	94
155	46	215 - 217	95
156	47	218 - 221	96
157	48	222 - 225	97
158	49	226 - 229	98
		230 - 265	99

Table A-8. AFOQT - P_2 Provisional Conversion Table for Academic Aptitude Composite

Raw	Percentile	Raw	Percentile
00 - 27	01	100	51
28 - 35	02	101	52
36 - 41	03	102	53
42 - 44	04	103	54
45 - 48	05	104	57
49 - 50	06	105	59
51 - 53	07	106	61
54	08	107	62
55 - 58	09	108	63
59 - 60	10	109	65
61	11	110	67
62	12	111	68
63 - 64	13	112	69
65	14	113	70
66	15	114	71
67 - 68	16	115	72
69	17	116	75
70 - 72	18	117	76
73	19	118	78
74	20	119	79
75	21	120	80
76	22	121	81
77	23	122	82
78	24	123	83
79	25	124	84
80	26	125	85
81	27	126	86
82	28	127	87
83 - 84	29	128	88
85	31	129	89
86	33	130	90
87	34	131	91
88	35	132	92
89	36	133 - 134	93
90	37	135	94
91	38	136 - 138	95
92	40	139 - 140	96
93	41	141 - 142	97
94	43	143 - 144	98
95	44	145 - 150	99
96	45		
97	47		
98	49		
99	50		

Table A-9. AFOQT - P_2 Provisional Conversion Table for Verbal Composite

Raw	Percentile	Raw	Percentile
00 - 14	01	45	40
15	02	46	41
16 - 18	03	47	44
19	04	48	46
20	05	49	48
21	06	50	50
22	07	51	53
23 - 24	08	52	55
25	09	53	57
26	10	54	60
27	11	55	62
28	12	56	67
29	13	57	69
30	14	58	72
31	15	59	74
32	17	60	77
33	18	61	78
34	19	62	81
35	21	63	84
36	23	64	86
37	24	65	87
38	26	66	90
39	27	67	92
40	30	68	93
41	32	69	96
42	33	70	97
43	36	71	98
44	38	72 - 75	99

Table A-10. AFOQT - P_2 Provisional Conversion Table for Quantitative Composite

Raw	Percentile	Raw	Percentile
00 - 12	01	50	52
13 - 16	02	51	54
17 - 19	03	52	57
20	04	53	59
21 - 22	05	54	61
23	06	55	64
24 - 25	08	56	66
26	09	57 - 58	69
27 - 28	10	59	71
29	11	60	75
30	14	61	76
31 - 32	15	62	78
33	17	63	80
34	19	64	82
35 - 36	21	65	85
37	24	66	86
38	26	67	88
39	28	68	90
40 - 41	31	69	91
42	33	70	92
43	34	71	93
44	38	72	95
45	41	73	96
46	43	74	97
47 - 48	45	75	98
49	48		